

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
14 June 2001 (14.06.2001)

PCT

(10) International Publication Number
WO 01/43328 A1

(51) International Patent Classification⁷: **H04J 14/08**

(21) International Application Number: **PCT/US00/33610**

(22) International Filing Date:
11 December 2000 (11.12.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/170,184 10 December 1999 (10.12.1999) US
60/183,479 18 February 2000 (18.02.2000) US

(71) Applicant (for all designated States except US): **QLOGIC SWITCH PRODUCTS, INC. [US/US]**; Suite 13, 6321 Bury Drive, Eden Prairie, MN 55346 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **GEORGE, William**,

R. [US/US]; 4744 Upton Avenue South, Minneapolis, MN 55410 (US). **BETKER, Steven, M.** [US/US]; 3196 Park Overlook, Shoreview, MN 55126 (US).

(74) Agent: **VIKSNINS, Ann, S.; Schwegman, Lundberg, Woessner & Kluth, P.O. Box 2938, Minneapolis, MN 55402 (US).**

(81) Designated States (national): CA, JP, US.

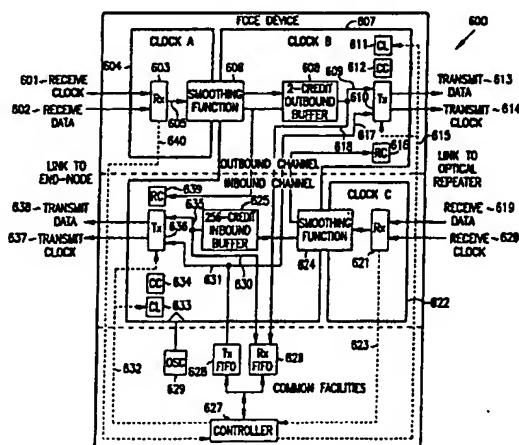
(84) Designated States (regional): European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:

— With international search report.

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: **FIBRE CHANNEL CREDIT EXTENDER AND REPEATER**



WO 01/43328 A1

(57) Abstract: The Fibre Channel Credit Extender (FCCE) (600) is a network device that is disposed between and connected to an end node (210) and an optical repeater (220). The FCCE (600) contains as many buffer credits as necessary, to solve bandwidth problems in a network. In a situation where maximum bandwidth is required in both directions of a link, the FCCE (600) breaks a single logical link into three physically separated "linklets". The short-distance linklets attain maximum bandwidth by use of the existing buffer credits of the end nodes. The long-distance linklet attains maximum bandwidth by use of very high receive buffer credits in the FCCEs (600). In this way, only those links that need maximum bandwidth over distances not covered by end-node credit counts need be attached to an FCCE (600). The FCCE (600) contains the optical repeater to gain distance on that link, and contains high credit count receive buffers to gain bandwidth on the link. All other ports of the switch can have smaller and less expensive receive buffers.

FIBRE CHANNEL CREDIT EXTENDER AND REPEATER

FIELD

5 This invention relates generally to optical communications and more particularly to a fibre channel network device.

BACKGROUND

10 The need for high performance switching solutions continues to grow in the fields of computing and data handling systems. Examples of such systems include interconnecting computers and high-performance storage devices, interconnecting computers in a multiple-computer operating environment, and anywhere else where multiple high-speed data interconnections must be established between designated nodes or groups of nodes in a data handling network. A switch is a network device at a node that sends and receives data 15 across the network in units of frames. Higher bandwidth and greater switching flexibility are prime concerns for switches and devices to be used in such systems.

20 The Fibre Channel standard, ANSI X3.T11, is intended to address these concerns. The Fibre Channel standard itself broadly defines classes and standards of performance, but does not dictate the implementation technologies to be used in providing these functions. A particular design of a switch to implement Fibre Channel functions is referred to as the "fabric" of the switch.

25 In order to increase the physical distance between switches, they often contain optical repeaters that transmit data across the network. The problem is that although the optical repeaters give the distance required, they almost always result in a very low sustainable bandwidth, especially if the link distance between nodes is quite long, such as 100 kilometers or more.

30 Thus, most conventional switches contain memory called buffers to hold the frames received and sent across the network. Associated with these buffers are credits, which are the number of frames that a buffer can hold per fabric port.

Most existing FC switches have approximately 8-32 credits per fabric port. These easily meet most requirements for longwave and shortwave links. Recently, the demand for longer links has increased, where 100 kilometer links are very popular. 100 kilometer links require approximately 62 credits per link
5 receiver at 1G, 124 credits at 2G, and 248 credits at 4G. It's not always possible, practical or desirable, to have available this much credit at the end of long links, especially FC switches, due to cost and integration concerns. Given especially that switch users would like to connect a long link to any switch port, it forces all switch ports to have a very large credit count or dynamic access to a very large credit count. This is especially impractical given that in larger switch fabrics consisting of multiple switch boxes, the E_Ports (or trunk ports) usually require very little buffering because of the short interconnect. Forcing large credit count buffers onto all ports of a switch increases cost and precludes highly integrated architectures.

15 In addition, the link has to be routed through optical repeaters on both ends, in order to operate reliably over the 100 kilometers. A typical installation has each end node connected to an optical repeater box, typically in the same room and typically via a shortwave cable. The long link then is actually between the two optical repeaters and what is typically commercially available dark fiber.

20 Thus, there is a need for a technique to increase the performance of switches.

SUMMARY

The present invention provides further improvements in high performance switching networks and methodology, and for providing a practical
25 implementation of Fibre Channel protocols.

In one aspect, the present invention achieves this through a credit extender. The Fibre Channel Credit Extender (FCCE) is a network device that is disposed between and connected to an end node and an optical repeater. The FCCE contains as many buffer credits as necessary, to solve the bandwidth

problem. In a situation where maximum bandwidth is required in both directions of a link, the FCCE breaks a single logical link into three physically separated "linklets." The short-distance linklets attain maximum bandwidth by use of the existing buffer credits of the end nodes. The long-distance linklet attains 5 maximum bandwidth by use of very high receive buffer credits in the FCCEs.

In this way, only those links that need maximum bandwidth over distances not covered by end-node credit counts need be attached to an FCCE. The FCCE contains the optical repeater to gain distance on that link, and contains high credit count receive buffers to gain bandwidth on the link. All other ports of 10 the switch can have smaller and cheaper receive buffers.

According to another aspect of the invention, the present invention provides a non-transparent fibre channel credit based repeater (FCBR) included with an end-node fabric manager, which controls FCBR initialization and distribution of buffer credits. Since, in this aspect, there is no need for the FCBR 15 to sink and source frames, frame buffers and an outbound link buffer are not needed in the FCBR, which simplifies the design.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 depicts a block diagram of a prior art fibre channel link.

Fig. 2 depicts a block diagram of a 100 KM link with inline FCCEs, 20 coupled together in the same box as the optical repeaters, according to an embodiment of the invention.

Fig. 3 depicts a block diagram of the FCCE installed on only one end, which is the end that is receiving the high bandwidth, according to an embodiment of the invention.

25 Fig. 4 depicts a block diagram of the duplex link simplified, showing only the credit mechanisms, which determine bandwidth on the link, according to an embodiment of the invention.

Fig. 5 depicts a block diagram of an example flowchart that describes a method for intercepting login request and response frames in the FCCEs for manipulating credit counts, according to an embodiment of the invention.

5 Fig. 6 depicts a block diagram of an example FCCE device, according to an embodiment of the invention.

Fig. 7 depicts an example flowchart for setting up credit for a single FCCE where the original login request frame was sourced at the left end node, according to an embodiment of the invention.

10 Fig. 8 depicts a block diagram of an example of a prior art fibre channel link with optical repeaters.

Fig. 9 depicts a block diagram of a 100 KM link with the inline FCBRs (Fibre Channel Credit Based Repeaters), coupled with optical repeaters, according to an embodiment of the invention.

15 Fig. 10 depicts a block diagram of a link with half-duplex FCBRs where the left-to-right direction has the maximum bandwidth, according to an embodiment of the invention.

Fig. 11 depicts a block diagram of the duplex link that is simplified showing only the credit mechanisms, which are what determines bandwidth on the link, according to an embodiment of the invention.

20 Fig. 12 depicts a block diagram that illustrates example non-transparent FCBR credit domains, according to an embodiment of the invention.

Fig. 13 depicts a block diagram that illustrates an example FCBR device suitable for use in the non-transparent environment, according to an embodiment of the invention.

DETAILED DESCRIPTION

In the following detailed description of exemplary embodiments of the invention, reference is made to the accompanying drawings (where like numbers

represent like elements), which form a part hereof, and in which is shown by way of illustration specific exemplary embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, but other embodiments may be utilized and logical, mechanical, electrical, and other changes may be made without departing from the scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims. For example, although embodiments of the present invention are described in the context of the fibre channel bus, in other embodiments, any suitable bus can be used.

10 The detailed description is divided into two embodiments, (I) the fibre channel credit extender embodiment and (II) the fibre channel credit based repeater embodiment.

I. FIBRE CHANNEL CREDIT EXTENDER EMBODIMENT

15 The following nine statements describe one embodiment of a fibre channel credit extender:

20 1. An FCCE (fibre channel credit extender) can be added onto any Fibre Channel non-Arbitrated Loop link utilizing R_RDY flow control and FC standard login functions, which includes N_Ports, F_Ports and E_Ports, to increase the maximum sustainable bandwidth where bandwidth would otherwise be lost due to insufficient end-node credit.

25 2. The FCCE is an inline link device that provides full bandwidth on any length link by inserting inline additional credits, up to the limit of 256 Fibre Channel credits.

25 100 MB/s up to 412 KM for 1G links with 256 credits

200 MB/s up to 206 KM for 2G links with 256 credits

400 MB/s up to 103 KM for 4G links with 256 credits

3. An FCCE can be installed at both ends of a link for duplex full bandwidth, or at one end of a link for full bandwidth in one direction only.
4. An FCCE is transparent to the end nodes. Each end node logs into the other end node, even though there are intermediate devices. All link initialization and reinitialization intended to cover the link, covers all linklets.
5. Credit assignment at FCCEs and end nodes can be accomplished either by explicit fabric management commands, or by FCCE transparent manipulation of the "Buffer-To-Buffer Credit" field of FLOGI (fabric login) and PLOGI (processor login) frames.
- 10 6. FC links are designed to achieve a maximum sustainable data transfer rate of approximately 100 MB/s if the link is running at a raw serial rate of 1.0625 Gb/s (1G), 200 MB/s if the link is running at 2.125 Gb/s (2g), or 400 MB/s if the link is running at 4.25 Gb/s (4G).
- 15 7. These maximum transfer rates can be achieved when FC adapters transmit long packets of data that are a sequence of back-to-back maximum length frames (2084 byte frames with 2048 byte payload), with a minimum of IDLE sequences (typically 6, but as many as 21) between each frame.
- 20 8. An additional requirement for maximum transfer rates is that there are sufficient receive buffer credits relative to the link length and raw transfer rate. If there are not enough receiver credits relative to link length and rate, actual sustainable bandwidth will be less than the maximum and can easily be much less than maximum, approaching less than 10% in some cases.
- 25 9. The general rule-of-thumb for calculating the number of receive credits required to achieve maximum bandwidth relative to length is 0.6 credits/kilometer for 1G, 1.2 credits/kilometer for 2G, and 2.4 credits/kilometer for 4G. Thus, a low number of credits is required to meet the standard Fibre Channel shortwave and longwave requirements. A 500 meter

link using 1G/2G/4G shortwave requires 0.3/0.6/1.2 credits, and 10 kilometer link using 1G/2G/4G longwave requires 6/12/24 credits.

5 Fig. 1 depicts an example prior art fibre channel link, which shows a 100 KM link utilizing only optical repeaters (Rep), and showing the receive buffers (Rb) whose credit counts determine the max link bandwidths.

The problem is that although the optical repeaters give the distance required, it almost always results in very low sustainable bandwidth, particularly if the link distance is quite long, especially 100 kilometers or more. The FCCE solves this problem as described below.

10 Fig. 2 shows a 100 KM link with the inline FCCEs, coupled together in the same box as the optical repeaters.

15 The Fibre Channel Credit Extender (FCCE) is a device that is between an end node 210 and an optical repeater 220 that contains as many buffer credits as necessary, to solve the bandwidth problem. In a situation where maximum bandwidth is required in both directions of a link, it essentially breaks a single logical link into three physically separated "linklets." The short-distance linklets (within the same room in one embodiment) attain maximum bandwidth by use of the existing buffer credits of the end nodes. The long-distance linklet (dark fiber up to 100 KM in one embodiment) attains maximum bandwidth by use of very high receive buffer credits in the FCCEs.

20 In this way, only those links that need maximum bandwidth over distances not covered by end-node credit counts, need be attached to an FCCE. The FCCE contains the optical repeater to gain distance on that link, and contains high credit count receive buffers to gain bandwidth on the link. All other ports of the switch can have smaller and cheaper receive buffers.

25 Fig. 3 shows the FCCE installed on only one end, which is the end that is receiving the high bandwidth. This embodiment provides a cost savings if the maximum bandwidth requirement is in one direction only. In Fig. 3, the left-to-

right direction has the maximum bandwidth. The same diagram would apply if duplex full bandwidth is required where one end node had sufficient credits, and the other end node did not.

5 Fig. 4 shows the duplex link simplified, showing only the credit mechanisms, which determine bandwidth on the link, as follows, where all receive buffers are shown with some suggested FCCE receive buffer credit values used in one embodiment that are high enough to solve the problem.

10 The end nodes show typical credit counts of 8 and 16. The FCCE has what may be ideal credit counts, where two credits are sufficient for the linklet length of 1-2 meters between the end node and the FCCE, and where 256 credits between the FCCEs on either end of the long linklet provides for maximum bandwidth on the longest distances possible, in one embodiment. However, 15 FCCE Receive Buffers can be designed with any credit count up to the maximum of 256, as per Fibre Channel rules.

15 Each linklet is a separate credit domain, and each follows Fibre Channel rules for R_RDY flow control within the domain. Shown in parentheses are the remote credit counts that each transmitter must deal with for the FCCE concept to work, in one embodiment. The FCCE provides a method where these credits are properly distributed prior to full bandwidth use of the link. In one embodiment, 20 these credits can be gained either by either of two methods: 1) explicitly programming each transmitter via fabric manager commands at all link devices, or 2) by intercepting the login request and response frames in the FCCEs for the purpose of manipulating the credit counts. In either case, all devices, including the end nodes and the FCCEs follow the Fibre Channel practice of assuming one 25 credit receive buffer until a larger credit count is specified.

The first method, i.e., explicit assignment of credits, is self-explanatory.

Fig. 5 depicts a flow chart for carrying out the second method, i.e., assignment of credits by FCCE login request/response frame interception and manipulation. All login frames are intercepted by the FCCE and delivered to the

5 fabric entity, which typically is a microprocessor, but which can be any form of a state machine. The FCCE fabric removes the credit count from the frame and installs it at the transmitter going back in the direction the frame arrived, then substitutes its own credit count in the login frame and forwards it in the same direction.

10 Control begins at block 500. Control then continues to block 510 where end node A forms a login request frame, containing all pertinent information, including its own credit count of 16 (in one embodiment), and transmits it. Control then continues to block 520 where FCCE B receives the login request and forwards it to the fabric entity. The fabric entity removes the credit count of 16 (in one embodiment) and programs it into the transmitter going in the right-to-left direction. The fabric entity replaces the credit count value of 16 (in one embodiment) with the value of its right-to-left receive buffer, 256 (in one embodiment), and forwards the frame in the left-to-right direction.

15 Control then continues to block 530 where FCCE C receives the login request and forwards it to the fabric entity. The fabric entity removes the credit count of 256 (in one embodiment) and programs it into the transmitter going in the right-to-left direction. The fabric entity replaces the credit count value of 256 (in one embodiment) with the value of its right-to-left receive buffer, 2 (in one embodiment), and forwards the frame in the left-to-right direction.

20 Control then continues to block 540 where end node D receives the login request and processes it by removing the credit count of 2 (in one embodiment) and programming it into the transmitter. End Node D forms the login reply frame, containing all pertinent information, including its own credit count of 8 (in one embodiment), and transmits it. Note that End Node D has received all login parameters from End Node A, except the credit count, which came from FCCE C.

25 Control then continues to block 550 where FCCE C receives the login reply frame and forwards it to the fabric entity. The fabric entity removes the credit count of 8 (in one embodiment) and programs it into the transmitter going

in the left-to-right direction. The fabric entity replaces the credit count value of 8 (in one embodiment) with the value of its left-to-right receive buffer, 256 (in one embodiment), and forwards the frame in the right-to-left.

Control then continues to block 560 where FCCE B receives the login
5 reply frame and forwards it to the fabric entity. The fabric entity removes the credit count of 256 (in one embodiment) and programs it into the transmitter going in the left-to-right direction. The fabric entity replaces the credit count value of 256 (in one embodiment) with the value of its left-to-right receive buffer, 2 (in one embodiment), and forwards the frame in the right-to-left
10 direction.

Control then continues to block 570 where End Node A receives the login
reply frame and forwards it to the fabric entity. The fabric entity removes the credit count of 2 (in one embodiment) and programs it into the transmitter. Note that End Node A has received all login parameters from End Node D, except the
15 credit count, which came from FCCE B.

Control then continues to block 599 where the process completes. When this process is complete, both end nodes have been involved in a compliant Fibre Channel login sequence with each other, exchanging all required parameters, the only exception being credit counts. This process is therefore transparent to the
20 end nodes.

This device can be implemented in either off-the-shelf devices such as FC Endecs and FPGAs in one embodiment, or as an ASIC specific to this function in another embodiment. In still another embodiment, the device can be implemented using a processor-based system where the processor reads and
25 executes instructions contained in memory. The instructions defining the functions of this embodiment can be delivered to the FCCE via a variety of signal-bearing media, which include, but are not limited to:

(1) information permanently stored on non-writeable storage media (e.g., read-only memory devices within a computer such as CD-ROM disks) readable by an unillustrated CD-ROM drive;

5 (2) alterable information stored on writeable storage media (e.g., floppy disks within a diskette drive, a tape within a tape drive, or disks within a hard-disk drive); or

(3) information conveyed by a communications media, such as through a computer or telephone network including wireless communications.

10 Such signal-bearing media, when carrying computer-readable instructions that direct the functions of the present invention, represent embodiments of the present invention.

Fig. 6 depicts an example FCCE device, according to an embodiment. The FCCE device interfaces to the end node on the left side, to the long distance link on the right side, and has three main functional areas.

15

- The Outbound Channel, sourcing data to the long-distance link.
- The Inbound Channel, sinking data from the long-distance link.
- The Common Facilities, which control the entire device.

20 The Outbound Channel

The Outbound Channel forwards end-node transmitted data to the optical repeater. The Outbound Channel exists primarily to provide a means to implement buffer-to-buffer flow control and as a means to intercept and manipulate credit fields in Login frames. The number of credits shown is very low because it is assumed that the distance between the FCCE and the end node is very short.

Receive Clock 601 and Receive Data 602 represent a parallel link interface, where the FCCE is either connected to an external serializer/deserializer, or directly connected to the end node in a typical glueless interconnect. The Receive Clock 601 becomes the logic clock for all logic in 5 Clock A (604).

The Receiver (Rx) function 603 processes incoming link data, providing the following functions.

- Converts 10B to 8B with error detection. If errors are detected on incoming frames, the frame EOF will be modified.
- Handles the primitive signal R_RDY by passing it to the transmitter credit state machine Counter CC 634.
- Handles the primitive sequences of LR, LRR, NOS, and LOS as per FC rules, and supplies these as status 640 to the Controller 627.
- Performs the Sync_Acquired, Sync-Lost state machines as per FC rules, supplying them as status 640 to the Controller 627.
- Parses frames, inspecting each for type and content.
- All incoming frames are written via path 605 into the Smoothing Function 606 using Clock A 604 timing.

The Smoothing Function 606 is used to smooth the frequency difference between Clock A 604 and Clock B 607. Typically both clock domains run at the identical nominal frequency, but which can vary by ± 100 ppm if in a FC 25 environment. The Smoothing function will add FC primitives if the write side clock is slower than the read side clock, or will delete primitives if the opposite relationship is true. The FC primitives used for smoothing are IDLEs, LR, LRR,

NOS or OLS, depending on current traffic. Smoothing would occur between frames but never during a frame.

Clock B 607 is the main fabric clock, clocking transmissions in both directions. Its frequency source is a local oscillator 629, which is typically a harmonic of the nominal FC link frequency being used.

The 2-Credit Outbound Buffer 608 provides the re-transmission path of frames originally transmitted by the local end node. This buffer is required because the Outbound Channel receiver is in a different credit domain than is the Outbound Channel transmitter. The number of receive credits is low because of the assumed very short link distance to the local end node. This buffer routes selected frames such as FLOGI and PLOGI via path 618 to the Rx FIFO 626 and then to the Controller 627 for processing. Rx FIFO 626 performs a transformation from a coded to an uncoded signal. All other frames are routed via 609 to the Outbound Transmitter 610.

15 The Outbound Transmitter 610 has the following functions.

- Selects the data to be transmitted next, which could be:
 - IDLEs, if nothing else is to be transmitted, or
 - R_RDYs, if R_RDY Counter (RC) (616) commands, or
 - Link frames, if they exist in the Outbound buffer 608, or
 - Controller frames, if they exist in the Tx FIFO 628.
- Keeps track of the number of R_RDYs to be transmitted via RC 616. RC 616 is an up/down counter, which increments by one each time the Inbound Buffer 625 is emptied of a frame, and decrements each time an R_RDY is transmitted. When RC is non-zero, it commands the transmitter to source R_RDYs.

5

- Keeps track of credit consumed via Credit Counter (CC) 612. CC 612 is an up/down counter that increments every time a frame is transmitted, and decrements every time an R_RDY is received on the Inbound Channel. When RC value = LC value, all available credit has been consumed, and transmission of frames must cease until additional R_RDYs are received.
- Keeps track of max remote credit via Credit Limit (CL) register 611. CL 611 is a register that holds the maximum credit at the remote end of the link, which is compared against CC 612 to determine if credit exists for transmission. CL is written by Controller 627, which writes the appropriate credit limit as gleaned from a PLOGI or FLOGI frame.
- Converts internal 8B codes to link 10B codes.
- Formats data to the external parallel interface.

10

The parallel interface to Transmit Data 613 and Transmit Clock 614 typically feed the serializer side of a serializer/deserializer, which in turn feeds the outbound optical repeater.

Inbound Channel

15

The Inbound Channel forwards optical repeater received data to the end node. The Inbound Channel exists primarily to provide the high receive buffer credit count to attain the maximum link bandwidth, but also to provide the means to implement buffer-to-buffer flow control and as a means to intercept and manipulate credit fields in Login frames.

20

The Inbound Channel is identical to the outbound channel in all respects, except that the Inbound Buffer 625 holds a very large number of credits, to a maximum of 256, which is used to gain full bandwidth on the long link. This is the feature that FCCE uses to increase long link bandwidth.

Inbound Channel find numbers 619-624, and 630-639 are identical in function to Outbound Channel find numbers 601-607, 609-618 and 640.

Common Facilities

5 Controller 627 is a state machine that controls the action of this device, and may be implemented entirely in hardware, or it may be an embedded or non-embedded micro-controller or microprocessor executing software. Its primary functions are:

- Initialize the device on powerup.
- Initialize the links, both inbound and outbound.
- 10 • Reads login frames out of Rx FIFO 626.
- Distribute the credit counts in the original frames, and insert new credit counts before re-transmitting the frames.
- Writes login frames into Tx FIFO 628.
- Handle link exceptions.

15 The Rx FIFO 626 is a frame buffer that holds frames (typically PLOGI and FLOGI) that the Inbound Buffer 625 and Outbound Buffer 608 choose to re-vector to Controller 627, which reads the frames as they appear in Rx FIFO. Rx FIFO 626 performs a transformation from a coded to an uncoded signal.

20 The Tx FIFO 628 is a frame buffer that holds frames (typically PLOGI and FLOGI) that are destined to be transmitted on the Outbound Tx 610 or the Inbound Tx 636. The Controller 627 writes the frames to Tx FIFO, and a transmitter reads the frame when ready. TX FIFO 628 performs a transformation from an uncoded to a coded signal.

Buffer-To-Buffer Credit Manipulation

Fig. 7 depicts a example flowchart for setting up credit for a single FCCE where the original login request frame was sourced at the left end node, according to an embodiment of the invention. Fig. 7 depicts a method that is a subset of the prior description of setting up credit across the entire link consisting of two end nodes and two FCCEs. The processing depicted in Fig. 7 covers only a single FCCE where the original login request frame was sourced at the left end node.

Control begins at block 700. Control then continues to block 710 where the login frame arrives at the receiver 603, passes through the Smoothing Function 606 unchanged, and lands in the 2-Credit Outbound Buffer 608. The Outbound Buffer recognizes that the frame is FLOGI or PLOGI and instead of transmitting the frame on the outbound Tx 610, routes it instead via path 618 to Rx FIFO 626, which signals "full" to Controller 627.

Control then continues to block 720 where controller 627 removes the "Buffer-To-Buffer Credit" value from the frame and writes it to Inbound CL register 633. The Inbound Tx 636 can now handle the full credit limit of the local end node.

Control then continues to block 730 where controller 627 replaces the frame "Buffer-To-Buffer Credit" value with the value of the Inbound Buffer 625, which is 256. Controller 627 writes this frame to Tx FIFO 628, which is then transmitted by Outbound Tx 610 when Tx is ready.

Control then continues to block 740 where eventually, the login acknowledge frame is received from the other end of the long link. If there is an FCCE at the other end, the "Buffer-To-Buffer Credit" is that of the remote FCCE Inbound Buffer. If not, the "Buffer-To-Buffer Credit" is that of the remote end-node receive buffer.

Control then continues to block 750 where the login frame arrives at the receiver 621, passes through the Smoothing Function 624) unchanged, and lands in the 256-Credit Inbound Buffer 625. The Inbound Buffer recognizes that the

frame is FLOGI or PLOGI and instead of transmitting the frame on the Inbound Tx 636, routes it instead via path 630) to Rx FIFO 626, which signals “full” to controller 627.

5 Control then continues to block 760 where controller 627 removes the “Buffer-To-Buffer Credit” value from the frame and writes it to Outbound CL register 611. The Outbound Tx 610 can now handle the full credit limit of the remote receive buffer.

10 Control then continues to block 770 where controller 627 replaces the frame “Buffer-To-Buffer Credit” value with the value of the Outbound Buffer 628, which is 2. Controller 627 writes this frame to Tx FIFO 628, which is then transmitted by Inbound Tx 636 when Tx is ready. Control then continues to block 799 where the function ends. The FCCE has now initialized credit and normal frame traffic can start.

FCCE Controller

15 The following are some of the Fibre Channel functions that controller 627 is involved in.

ELS Frame Processing

Some Extended Link Service (ELS) frames need to be modified. These are frames where buffer-to-buffer flow control data has to be manipulated.

FLOGI (Fabric Login)

20 The FCCE starts out with Fabric mode off. If an FLOGI frame is received, the Fabric mode flag is turned on and the “Buffer-To-Buffer Credit” field in the FLOGI frame is changed to manipulated as described earlier. The OX_ID is remembered so that the ACC reply 25 can be identified.

PLOGI (N-Port Login)

If the FCCE is not in Fabric mode, the “Buffer-To-Buffer Credit” field in the PLOGI is manipulated as described earlier. The OX_ID is remembered so that the ACC can be identified.

5

FDISC

If the FCCE is in Fabric mode, the “Buffer-To-Buffer Credit” field in the FDISC frame is manipulated as described earlier. The OX_ID is remembered so that the ACC can be identified.

PDISC

10

If the FCCE is not in Fabric mode, the “Buffer-To-Buffer Credit” field in the PDISC frame is manipulated as described earlier. The OX_ID is remembered so that the ACC can be identified.

ACC (ELS Accept)

15

If the ACC OX_ID matches a remembered ELS OX_ID and the frame is at least the minimum length for a login ACC (116 bytes), the frame is assumed to be a reply to a login frame. The “Buffer-To-Buffer Credit” field is manipulated as described earlier. The remembered OX_ID is cleared.

SW ILS Frame Processing

20

These frames are used to configure Fabrics created by linking multiple switches together.

ELP (Exchange Link Parameters)

25

The “Buffer-To-Buffer Credit” field in the ELP is manipulated as described earlier. The OX_ID is remembered so that the SW_ACC can be identified. If ELP frames have been received in both directions, the FCCE goes into Fabric mode.

SW_ACC (Switch Services Accept)

5

If the SW_ACC OX_ID matches a remembered ELP OX_ID, and the frame is the minimum length for an ELP SW_ACC, the frame is assumed to be a reply to the ELP. The "Buffer-To-Buffer Credit" field is manipulated as described earlier. The remembered OX_ID is cleared. The FCCE goes into Fabric mode.

Primitive Processing

The FCCE will generally pass received primitives on to the other end, or discard them (AL primitives).

10

NOS, OLS, LR, LRR Primitives

15

If one of these primitives is received, the FCCE may send or discard frames in the buffers for the channel. No R_RDYs are sent for frames that are transmitted out the other side while a primitive is being received. Once all received frames are gone, the currently received primitive is sent to the channel transmitter. Flow control is reset by clearing the CC counter for the channel receiving the primitive and the RC counter of the other channel.

20

AL Primitives

All Arbitrated Loop primitives (LIP, ARB, etc.) are discarded and replaced with the Idle ordered set.

25

Loss of Sync

If sync is lost for R_T_TOV (100 milliseconds), all frames in the channel buffer are discarded. The channel transmitter transmits a 10 bit code of all 0, which should cause loss of sync at the other end of the link.

II. FIBRE CHANNEL CREDIT BASED REPEATER EMBODIMENT

5 In the fibre channel credit based repeater embodiment, a FCBR (Fibre Channel Credit Based Repeater) can either be transparent to the end nodes, or non-transparent.

10 For a transparent application, the end node fabric management facilities are not aware of the FCBR presence, and do not contain FCBR-specific functions used for FCBR management. FCBR initialization, error handling and assignment of "BB_Credit" are FCBR functions.

15 10 For a non-transparent application, the end node fabric management facilities are aware of the FCBR presence, and contain FCBR-specific function used for FCBR management. FCBR initialization, error handling and assignment of "BB_Credit" are end-node functions.

15 Fig. 8 depicts a block diagram of a prior art example of a fibre channel link with optical repeaters, which shows a 100 KM link utilizing only optical repeaters (Rep), and showing the receive buffers (Rb) whose credit counts determine the max link bandwidths.

20 When the End Node receive buffers Ra and Rb contain typical credit counts of 8-32, the optical repeaters give the 100 KM distance required, but the maximum bandwidth on either link will be 12% (at 8 credits) to 50% (at 32 credits) of the theoretical maximum for 100 MB/s at 1G. The FCBR solves this problem.

The Transparent FCBR

25 The Fibre Channel Credit Based Repeater (FCBR) is a device that is connected between an end node and an optical repeater, that contains as many buffer credits as necessary, to solve the bandwidth problem. In a situation where maximum bandwidth is required in both directions of a link, it essentially breaks a single logical link into three physically separated "linklets." The short-distance

linklets (typically within the same room) attain maximum bandwidth by use of the existing buffer credits of the end nodes. The long-distance linklet (typically dark fiber up to 100 KM) attains maximum bandwidth by use of very high receive buffer credits in the FCBRs.

5 Fig. 9 depicts a block diagram of a 100 KM link with the inline FCBRs, coupled with optical repeaters, according to an embodiment of the invention.

Buffers Re and Rd are at the end of the long links, and contain enough credits (approximately 60 credits for 100 KM at 1 Gb in one embodiment) to insure maximum bandwidth on the link. Buffers Rc and Rf are very small (approximately 2 credits for a connection within the same room, in one embodiment) and exist purely to maintain the linklet credit management. In this way, only those links that need maximum bandwidth over distances not covered by end-node credit counts, need be attached to an FCBR. The FCBR shown contains the high credit count receive buffers to gain bandwidth on the link. All other ports of a switch can have smaller and cheaper receive buffers.

20 As a cost savings, if the maximum bandwidth requirement is in one direction only, the FCBR can be installed on only one end, the end that is receiving the high bandwidth. In Fig. 10, the left-to-right direction has the maximum bandwidth. The same diagram also applies when duplex full bandwidth is required, where Ra has sufficient credits, and Rb does not.

25 Fig. 11 depicts a block diagram of the duplex link that is simplified showing only the credit mechanisms, which are what determines bandwidth on the link, where all receive buffers are shown with some suggested FCBR receive buffer credit values high enough to solve typical bandwidth problems.

The end nodes show typical credit counts of 8 and 16. The FCBR has ideal credit counts for 100 Km at 1G, where 2 credits are sufficient for the linklet length of 1-2 meters between the end node and the FCBR, and 64 credits between the FCBRs on either end of the long linklet.

Each linklet is a separate credit domain, and each follows Fibre Channel rules for R_RDY flow control within the domain. Shown in parenthesis are the remote credit counts that each transmitter must deal with for the transparent FCBR concept to work. There must be a method where these credits are properly distributed prior to full bandwidth use of the link. These credits can be gained by intercepting the login request and response frames in the FCBRs for the purpose of manipulating the credit counts. Prior to BB_Credit disbursement by this method, all devices, including the end nodes and the FCBRs follow the Fibre Channel practice of assuming one credit receive buffers until a larger credit count is specified.

The processing for all login frames that are intercepted by the FCBR and delivered to the FCBR "fabric" is the same as that described above with reference to Fig. 5, except that the value of the right-to-left receive buffer in blocks 520 and 550 is 64 and the credit count in blocks 530 and 560 is 64. When this processing is complete, both end nodes have been involved in a compliant Fibre Channel login sequence with each other, exchanging all required parameters, the only exception being credit counts. This process is therefore transparent to the end nodes.

Transparent FCBR Hardware

The FCBR hardware can be the same as that previously described above with reference to Fig. 6.

Transparent FCBR Credit Manipulation

The buffer-to-buffer credit manipulation can be the same as that previously described above with reference to Fig. 7, except that in block 750 the credit inbound buffer has 64 credits instead of 256.

Primitive Processing

The primitive processing is the same as that previously described above with respect to the fibre channel credit extender embodiment, except for NOS, OLS, LR, and LRR primitives:

NOS, OLS, LR, LRR Primitives

5 If one of these primitives is received, the FCBR may send or discard frames in the buffers for the channel. No R_RDYs are sent for frames that are transmitted out the other side while a primitive is being received. Once all received frames are gone, the currently received primitive is sent to the channel transmitter. Flow control is reset by 10 clearing the CC counter for the channel receiving the primitive and the RC counter of the other channel.

The Non-Transparent FCBR Embodiment

15 This embodiment is much simpler than the Transparent FCBR embodiment, because in the non-transparent FCBR embodiment, the FCBR is in the realm of the end-node fabric manager facility, which controls FCBR initialization and distribution of BB_Credit. Since there is no need for the FCBR to sink and source frames, these buffers and the outbound link buffer are not needed.

20 Fig. 12 is a block diagram that depicts the non-transparent FCBR credit domains. Three differences relative to the transparent model are shown.

- The credit domains are reduced from 3 to 2.
- The FCBR fabric management is controlled by the end-node fabric managers.
- The long link advertised credit is the sum of the end-node credits and the FCBR credits.

25 When End Node A receive buffer is emptied of a frame, the resulting R_RDY is transmitted to FCBR B transmit state machine, and to End Node D

transmit state machine, bypassing the FCBR C buffer. Similarly, when End Node B receive buffer is emptied of a frame, the resulting R_RDY is transmitted to FCBR A transmit state machine, and to End Node A transmit state machine, bypassing the FCBR B buffer.

5 FLOGI frames are not intercepted by the FCBRs, but are passed unchanged to the end-node fabric managers, which program the correct BB_Credit values in both the end-node and the attached FCBR transmit state machines.

10 In the right-to-left path, the advertised credit available to the right end node is $64 + 16 = 80$. If the FCBR B receive buffer physically contains room for exactly 64 credits and no more, then the advertised credit is $64 + 16 - 1 = 79$. Holding back a single credit is required to prevent overflow at FCBR B when End Node D, for example, transmits a frame in response to an R_RDY from End Node A, but the frame hole has not migrated to FCBR B before End Node D
15 transmits the frame. Similarly, in the left-to-right path, the advertised credit can be $64 + 8 - 1 = 71$.

Fig. 13 is a block diagram that illustrates example FCBR device suitable for use in the non-transparent environment. Note the following differences from the transparent environment:

20

- There is no frame buffer to the long link, nor fabric manager frame sink and source facilities.
- All transmissions from the End Node are repeated onto the long link, including R_RDYs.
- R_RDYs from the local End Node are fed both to the long link and to the transmit state machine feeding the local End Node.
- R_RDYs from the remote end node are neither sunk nor used in the FCBR, but are passed onto the local End Node.

25

All local end Node transmissions are repeated onto the long link, subject to smoothing. Note that if the local end node device and the FCBR device derive clocks from the same oscillator, then the smoothing function can be deleted.

WHAT IS CLAIMED IS:

1. A method, comprising:
 - receiving a login request frame at a first network device from a first end node and forwarding the login request frame to a first fabric entity;
 - 5 programming a credit count contained in the login request frame into a first transmitter;
 - replacing the credit count in the login request frame with a value of a first receive buffer; and
 - transmitting the login request frame.
- 10 2. The method of claim 1, further comprising:
 - receiving the login request frame at a second network device and forwarding it to a second fabric entity;
 - programming the credit count contained in the login request frame into a second transmitter;
 - 15 replacing the credit count in the login request frame with a value of a second receive buffer; and
 - transmitting the login request frame.
- 20 3. The method of claim 2, further comprising:
 - receiving the login request frame at a second end node and programming the credit count contained in the login request frame into a third transmitter;
 - creating a login reply frame, wherein the reply frame includes a credit count of the end node; and
 - 25 transmitting the reply frame.
4. The method of claim 3, further comprising:
 - receiving the login reply frame at the second network device and forwarding the login reply frame to the second fabric entity;
 - 30 programming the credit count contained in the login request frame into the second transmitter;

replacing the credit count in the login reply frame with a value of the second receive buffer; and
transmitting the login request frame.

5 5. The method of claim 4, further comprising:
 receiving the login reply frame at the first network device and forwarding the login reply frame to the first fabric entity;
 programming the credit count contained in the login reply frame into the first transmitter;
10 10. replacing the credit count in the login reply frame with a value of the first receive buffer; and
 transmitting the login reply frame.

15 6. The method of claim 5, further comprising:
 receiving the login reply frame at the first end node and forwarding the login reply frame to the first fabric entity;
 programming the credit count contained in the login reply frame into an end node transmitter; and
 transmitting the login reply frame.

20 7. An apparatus, comprising:
 a first end node that forms a login request frame, wherein the login request frame comprises a credit count of the first end node, and transmits the login request frame;
25 25. a first network device that receives the login request frame, forwards the login request frame to a first fabric entity;
 wherein the first fabric entity programs the credit count contained in the login request frame into a first transmitter, and replaces the credit count in the login request frame with a value of a first receive buffer, and transmits the login request frame.

8. The apparatus of claim 7, further comprising:

5 a second network device that receives and forwards the login request frame to a second fabric entity, wherein the second fabric entity programs the credit count contained in the login request frame into a second transmitter, replaces the credit count in the login request frame with a value of a second receive buffer, and transmits the login request frame.

9. The apparatus of claim 8, further comprising:

10 a second end node that receives the login request frame and programs the credit count contained in the login request frame into a second-end-node transmitter, creates a login reply frame, wherein the reply frame includes a credit count of the end node, and transmits the reply frame.

15 10. The apparatus of claim 9, further wherein the second network device further:

20 receives the login reply frame and forwarding the login reply frame to the second fabric entity, wherein the second fabric entity programs the credit count contained in the login request frame into the second transmitter, replaces the credit count in the login reply frame with a value of the second receive buffer, and transmits the login request frame.

11. The apparatus of claim 10, wherein the first network device further:

25 receives the login reply frame and forwards the login reply frame to the first fabric entity, wherein the first fabric entity further programs the credit count contained in the login reply frame into the first transmitter, replaces the credit count in the login reply frame with a value of the first receive buffer, and transmits the login reply frame.

12. The apparatus of claim 11, wherein the first end node further:

30 receives the login reply frame and forwards the login reply frame to the first fabric entity, wherein the first fabric entity further programs the credit count

contained in the login reply frame into an end node transmitter and transmits the login reply frame.

13. A network device, comprising:

5 a controller that

reads a buffer-to-buffer credit value from a login frame and writes the buffer-to-buffer credit value to an inbound credit limit register, replaces the buffer-to-buffer credit value in the login frame with a value of an inbound buffer, and

10 writes the login frame to a first frame buffer, wherein the first frame buffer performs a transformation from an uncoded to a coded signal.

14. The network device of claim 13, further comprising:

15 an outbound buffer that recognizes that the login frame is FLOGI or PLOGI and routes the login frame to a second frame buffer that performs a transformation from a coded to an uncoded signal, wherein the second frame buffer sends a full signal to the controller

20 15. The network device of claim 14, further comprising:

a receiver that receives a login frame and passes the login frame to the outbound buffer.

16. A network device, comprising:

25 a controller that

reads a buffer-to-buffer credit value from the a login frame and writes the buffer-to-buffer credit value to an outbound credit limit register,

30 replaces the buffer-to-buffer credit value in the login frame with a value of an outbound buffer, and

writes the login frame to a first frame buffer that performs a transformation from an uncoded to a coded signal.

17. The network device of claim 16, further comprising:
5 an inbound buffer that recognizes that the login frame is FLOGI or PLOGI and routes the login frame to a second frame buffer performs a transformation from a coded to an uncoded signal, wherein the second frame buffer sends a full signal to the controller.

10 18. The network device of claim 17, further comprising:
a receiver that receives a login frame and passes the login frame to the inbound buffer.

15 19. A signal-bearing media comprising instructions, wherein the instructions when read and executed by a processor comprise:
receiving a login request frame at a network device from a first end node and forwarding the login request frame to a first fabric entity;
programming a credit count contained in the login request frame into a first transmitter;
20 replacing the credit count in the login request frame with a value of a first receive buffer; and
transmitting the login request frame.

25 20. The signal-bearing media of claim 19, further comprising:
receiving the login request frame at a second network device and forwarding it to a second fabric entity;
programming the credit count contained in the login request frame into a second transmitter;
replacing the credit count in the login request frame with a value of a second receive buffer; and
30 transmitting the login request frame.

21. The signal-bearing media of claim 20, further comprising:

receiving the login request frame at a second end node and programming the credit count contained in the login request frame into a transmitter;
5 creating a login reply frame, wherein the reply frame includes a credit count of the end node; and
transmitting the reply frame.

22. The signal-bearing media of claim 21, further comprising:

receiving the login reply frame at the second network device and
10 forwarding the login reply frame to the second fabric entity;
programming the credit count contained in the login request frame into a second transmitter;
replacing the credit count in the login reply frame with a value of a second receive buffer; and
15 transmitting the login request frame.

23. The signal-bearing media of claim 22, further comprising:

receiving the login reply frame at the first network device and forwarding the login reply frame to the first fabric entity;
20 programming the credit count contained in the login reply frame into the first transmitter;
replacing the credit count in the login reply frame with a value of the first receive buffer; and
transmitting the login reply frame.
25

24. The signal-bearing media of claim 23, further comprising:

receiving the login reply frame at the first end node and forwarding the login reply frame to the first fabric entity;
30 programming the credit count contained in the login reply frame into an end node transmitter; and
transmitting the login reply frame.

1/9

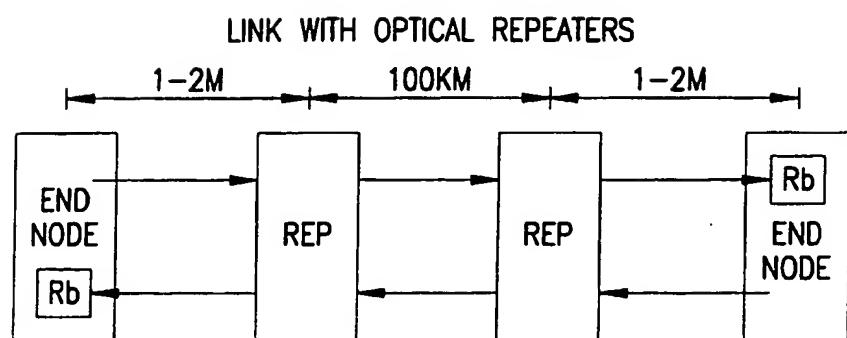


FIG. 1

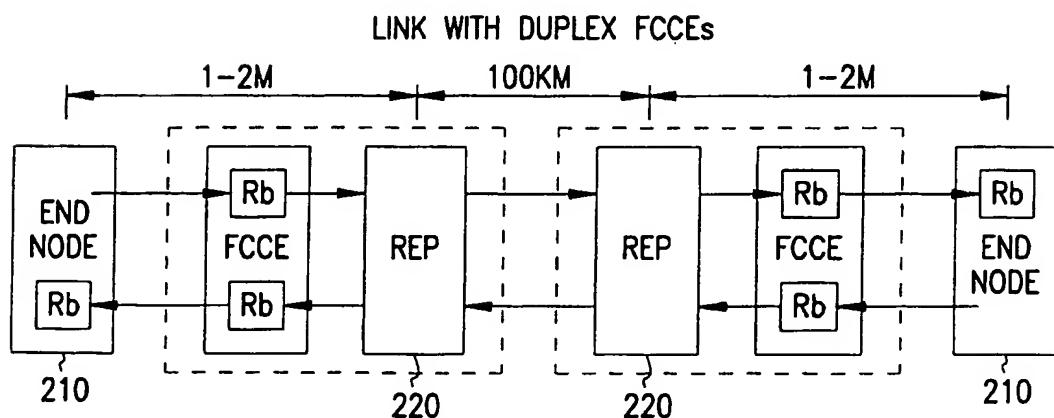


FIG. 2

2/9

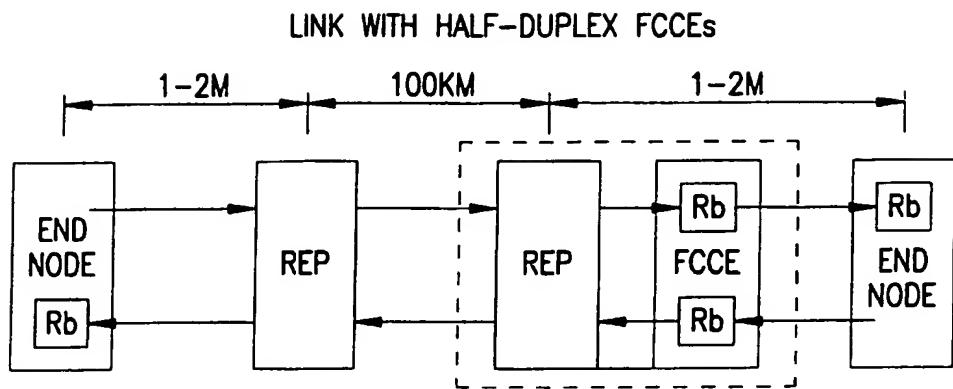


FIG. 3

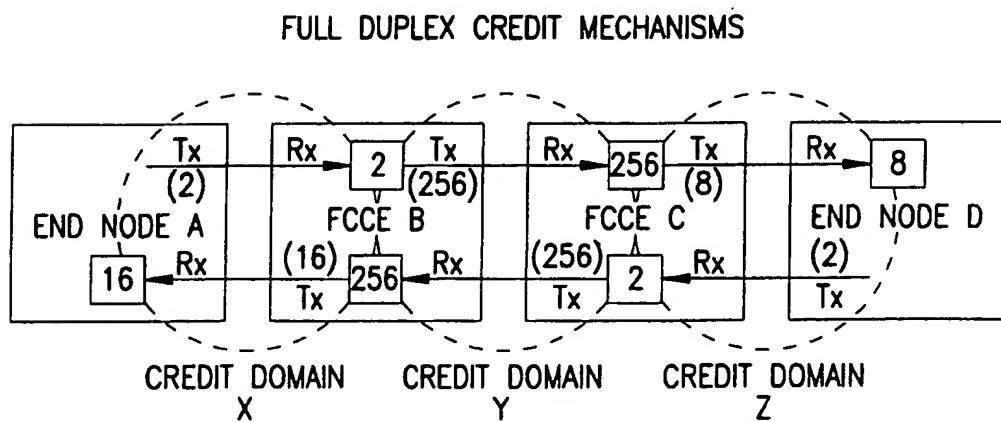


FIG. 4

3/9

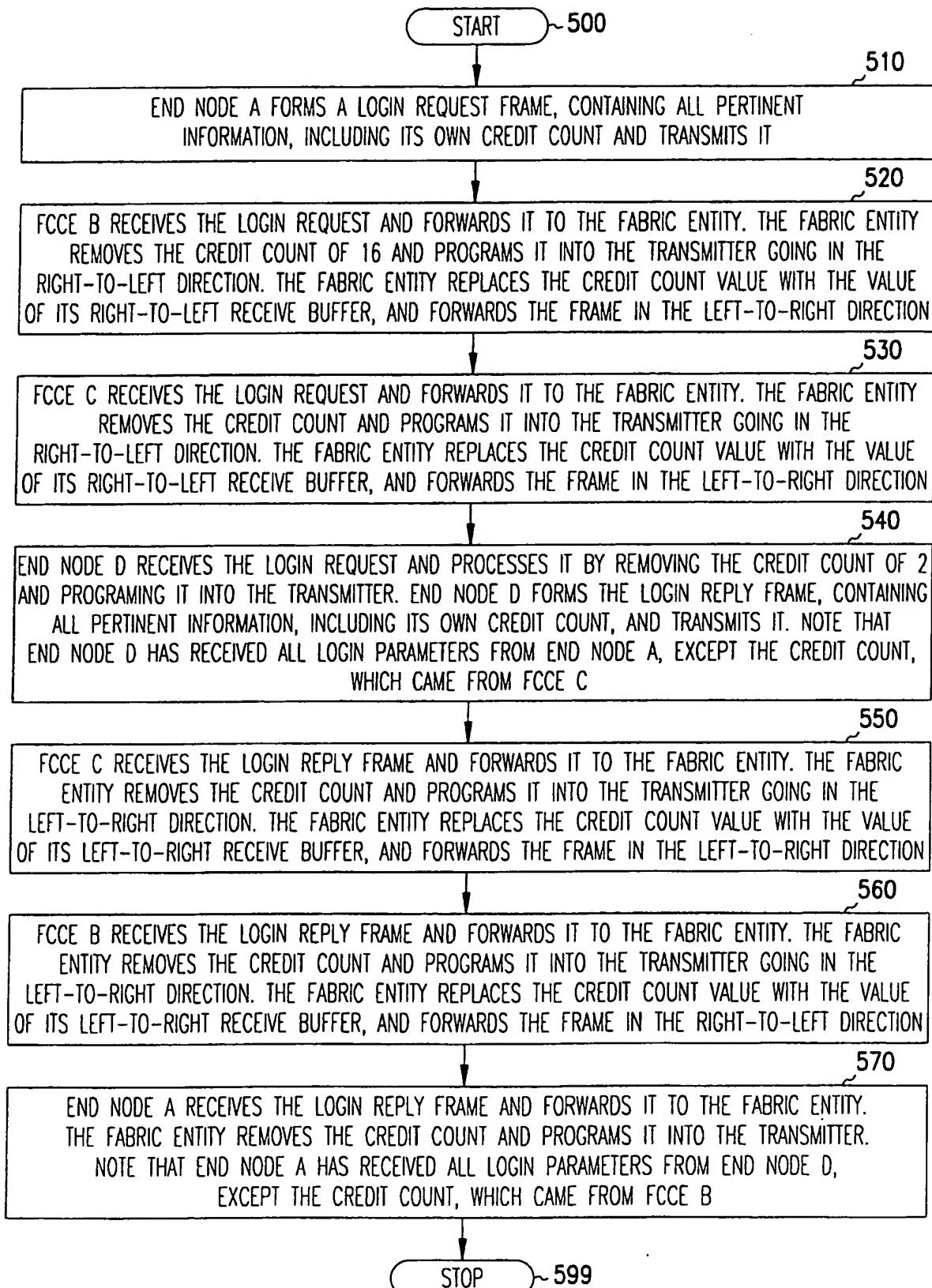


FIG. 5

4/9

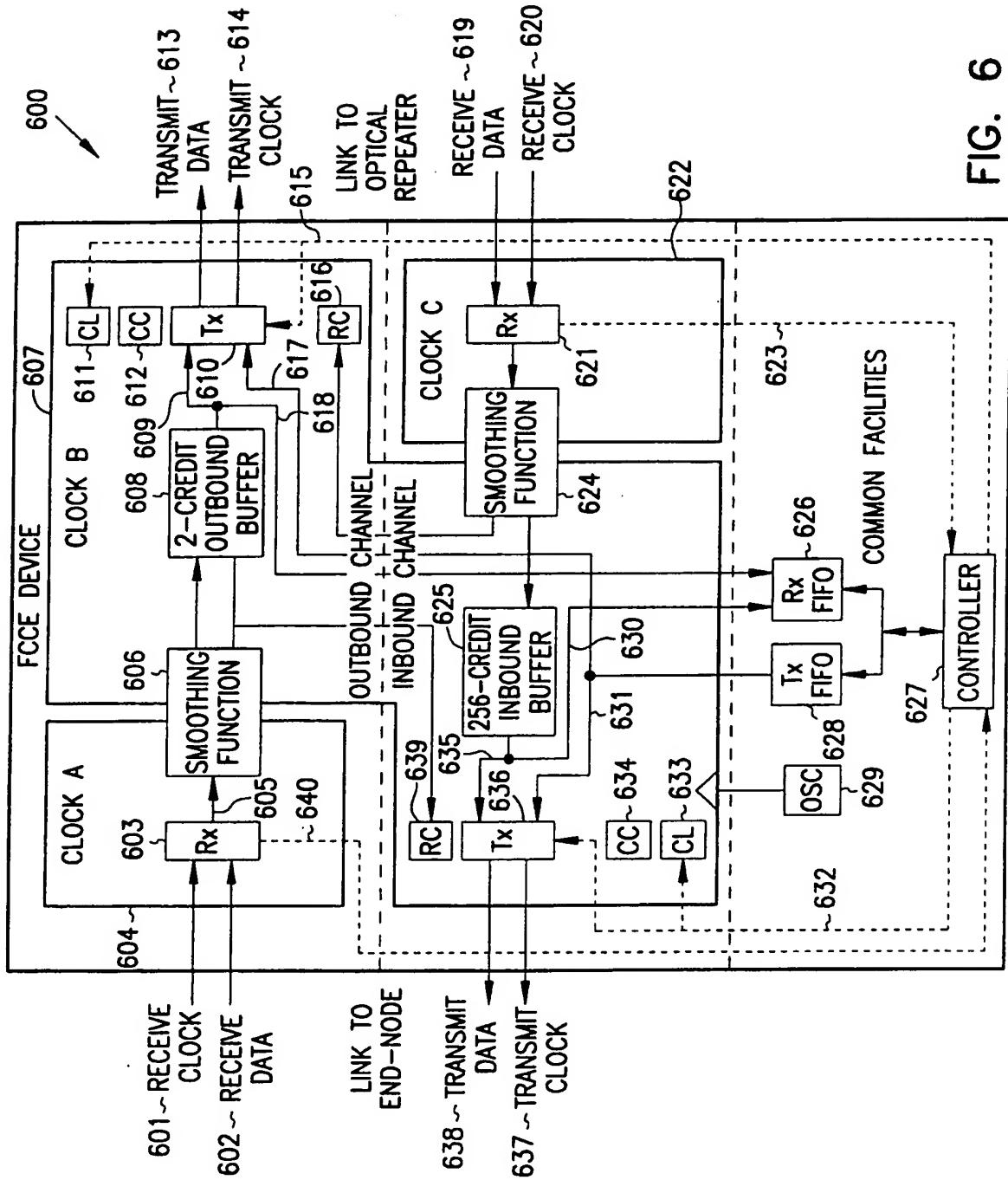


FIG. 6

5/9

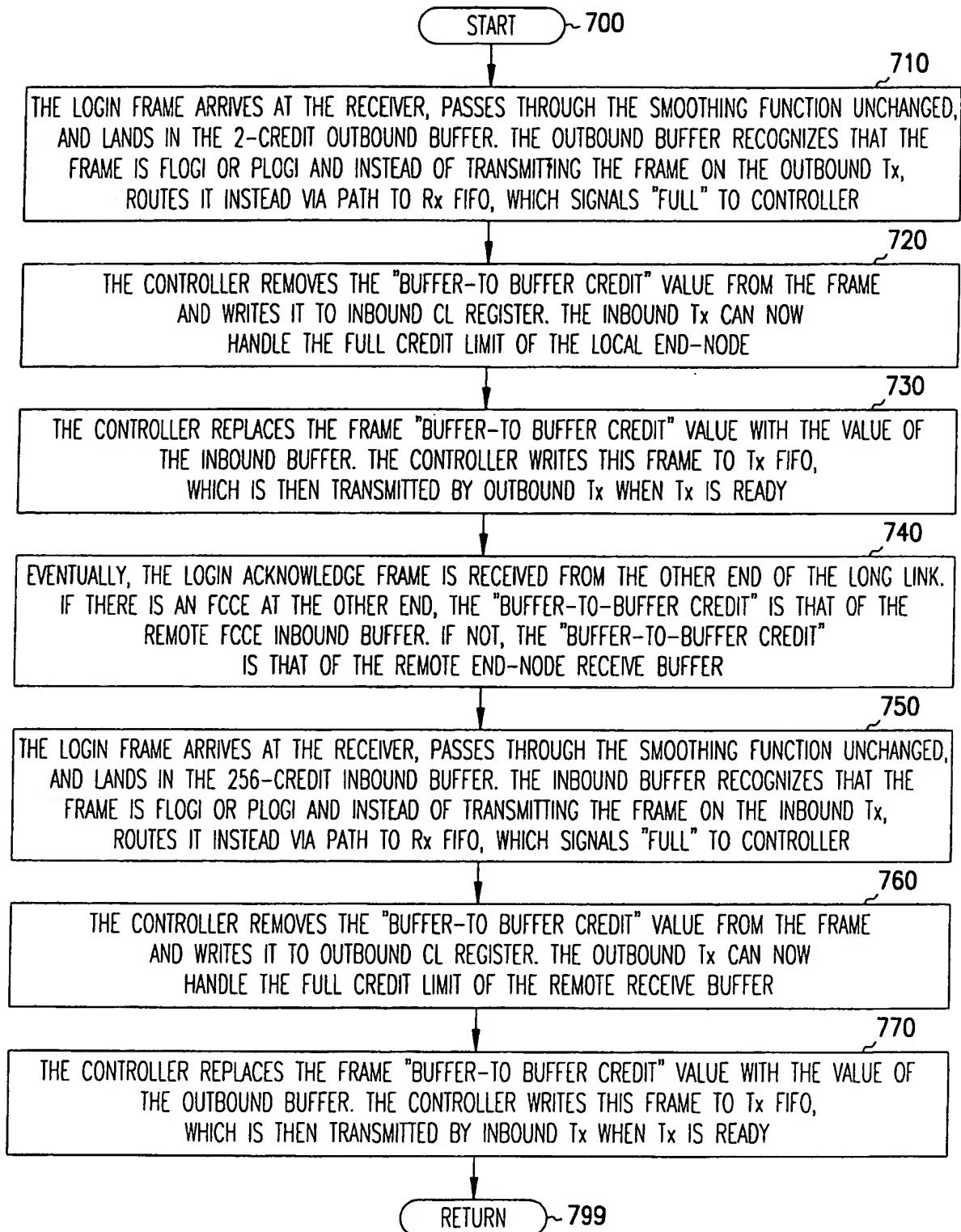


FIG. 7

6/9

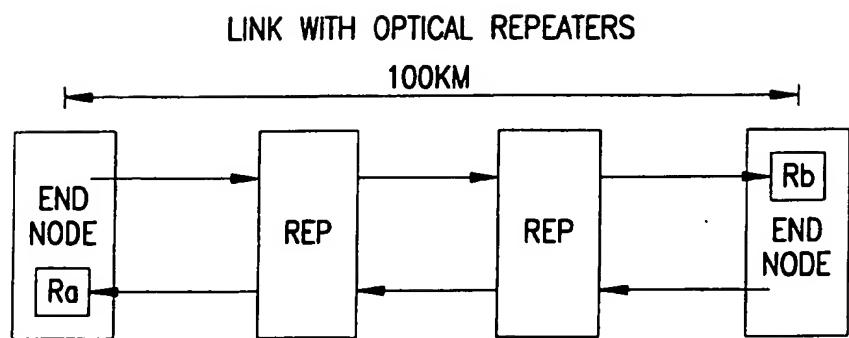


FIG. 8

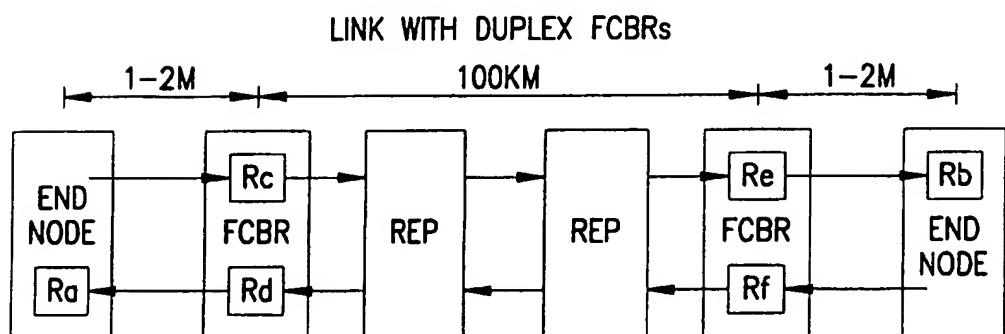


FIG. 9

7/9

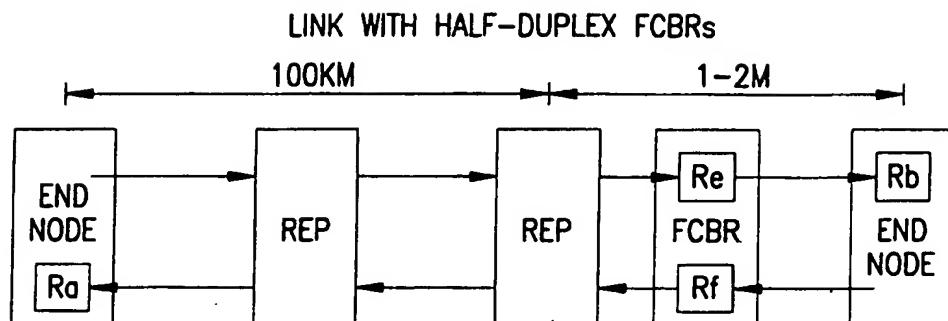


FIG. 10

FULL DUPLEX CREDIT MECHANISMS

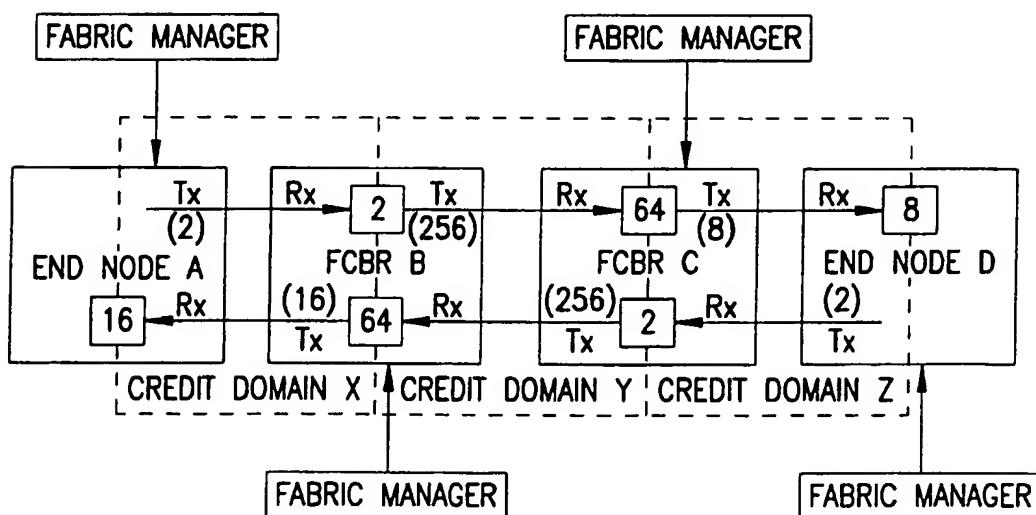


FIG. 11

8/9

NON-TRANSPARENT FCBR CREDIT DOMAINS

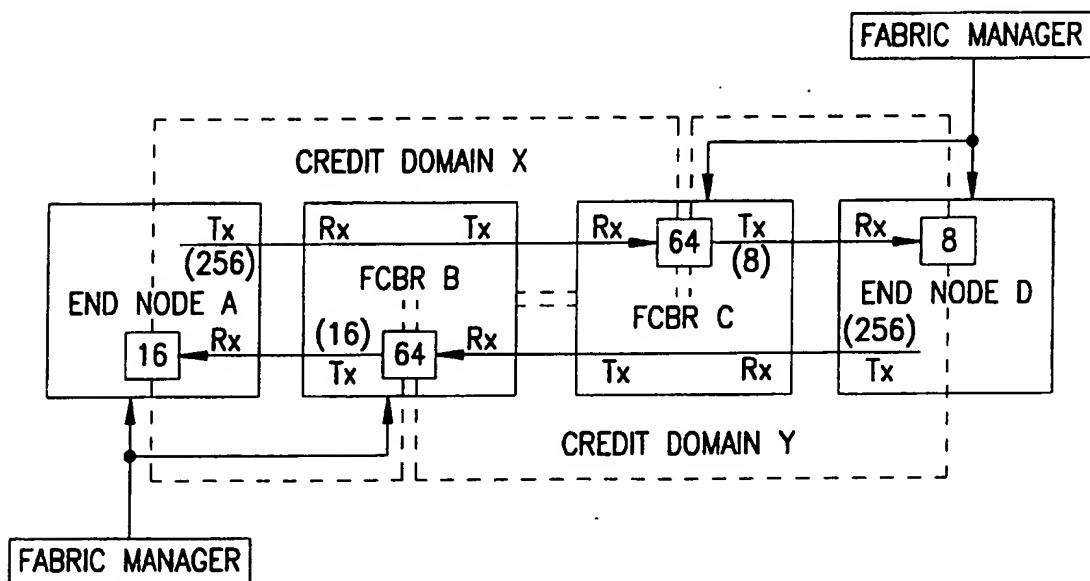
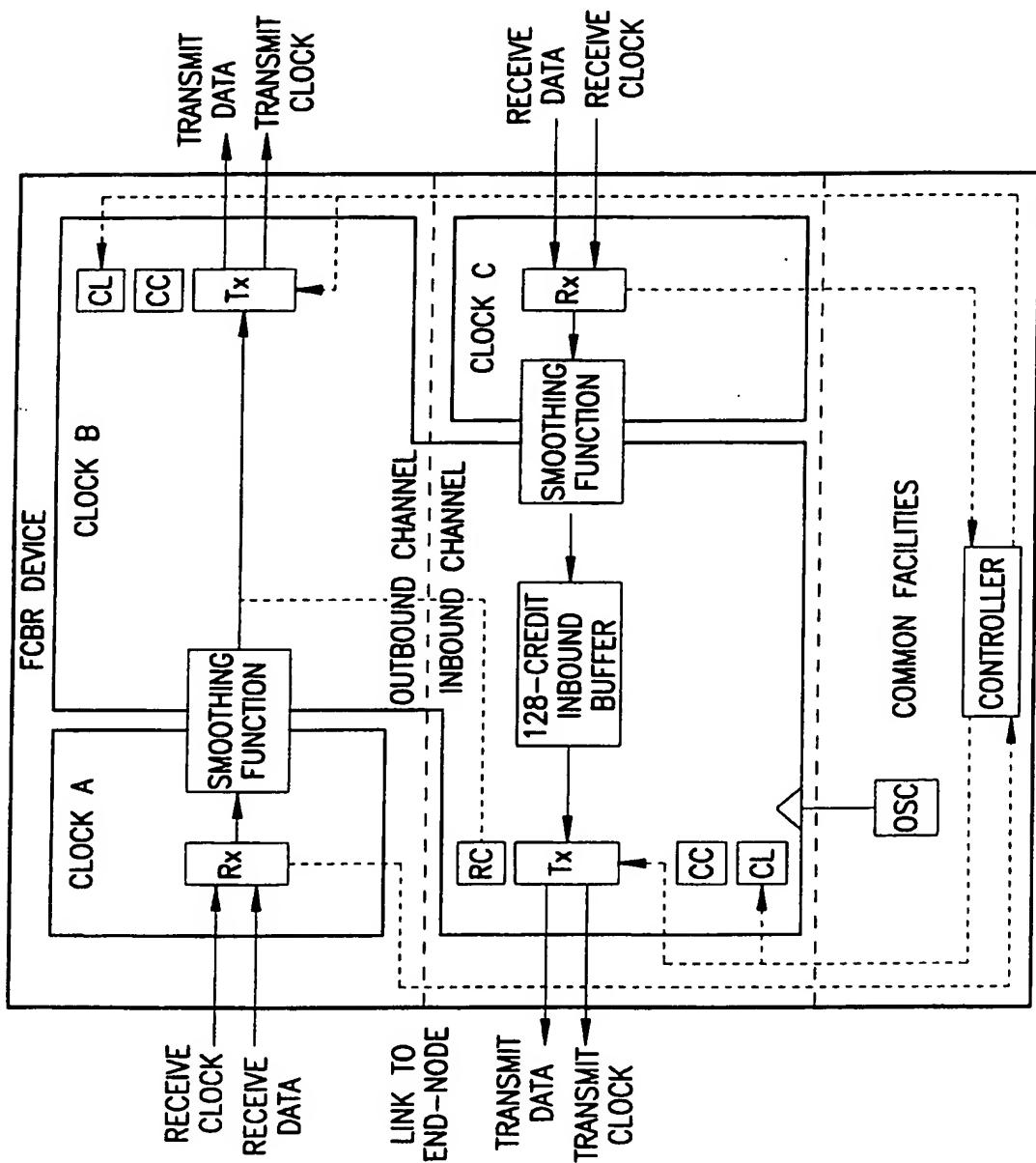


FIG. 12

9/9



13
FIG.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US00/33610

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : H04J 14/08
US CL : 370/235

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 370/232, 235, 236, 453, 906

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y,P	US 6,014,383 A (MCCARTY) 11 January 2000 (11.01.2000), column 7, lines 11-26.	1-24
Y	US 5,610,745 A (BENNETT) 11 March 1997 (11.03.1997), abstract.	1-24
Y	US 5,598,541 A (MALLADI) 28 January 1997 (28.01.1997), column 12, lines 45-66, column 16, lines 62-68.	1-24

Further documents are listed in the continuation of Box C.

See patent family annex.

• Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means		
"P" document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

16 February 2001 (16.02.2001)

Date of mailing of the international search report

30 MAR 2001

Name and mailing address of the ISA/US

Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Faxsimile No. (703)305-3230

Authorized officer

Chau Nguyen

Telephone No. (703)305-3906

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US00/33610

Continuation of B. FIELDS SEARCHED Item 3: EAST, IEEE
search terms: fibre channel, credit, login, fabric, buffer-to-buffer, extended link, forward, repeater